Линейная регрессия

План

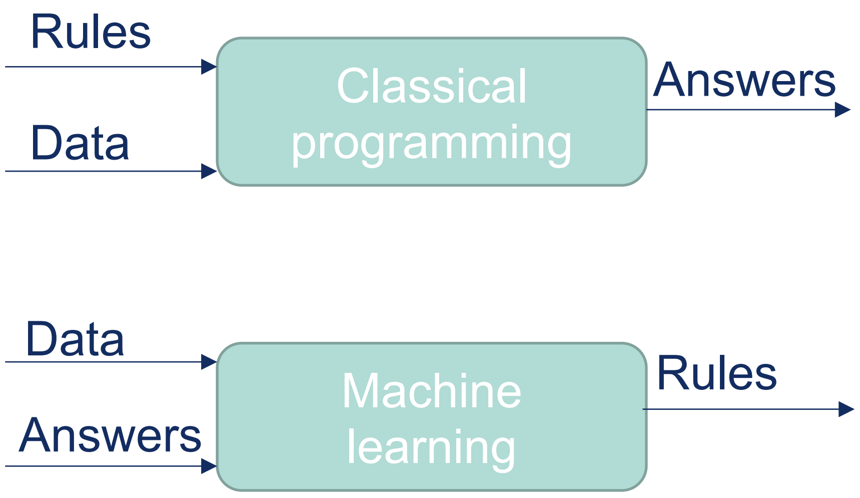
1. Введение
2. Линейная регрессия
3. Валидация моделей

Повторение

Что такое машинное обучение?

Машинное обучение – набор способов воспроизведения связей между событиями
И результатом

Машинное обучение – обширный подраздел искусственного интеллекта, изучающий методов
Построения алгоритмов, способных обучаться



Обозначения

- X– множество обьектов (характеристики ресторанов)
- Y – множество ответов, целевая переменная, target (прибыль от каждого ресторана)
- $a : X \to Y$ – неизвестная зависимость
- Какой из вариантов принесет максимальную прибыль?

Обучающая выборка

Дано:

- $\{x_1, \dots, x_n\} \subset X$ – обучающая выборка
- $\{y_1, \dots, y_n\}, y_i = y(x_i)$ – известные ответы

Найти:

- $a : X \to Y$ – алгоритм (решающая функция), приближающая

y на всем множестве X

Определения

- **Признаки, факторы (features)** - количественные характеристики обьекта

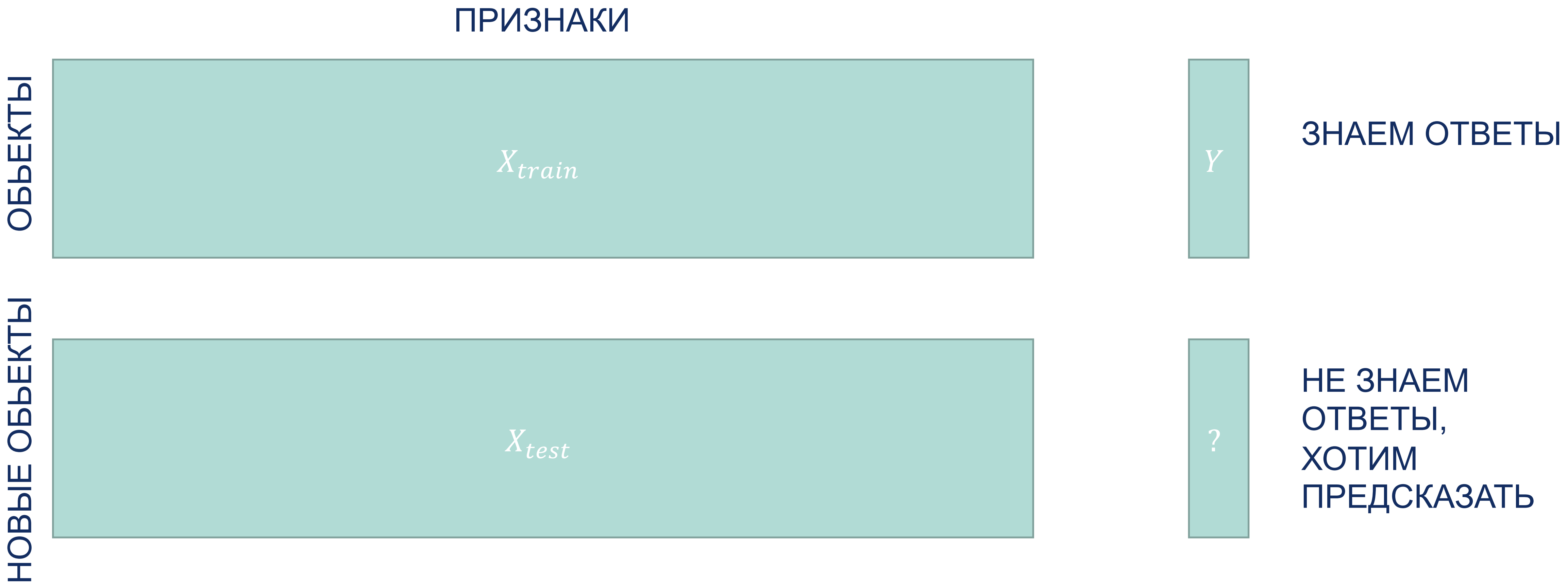
- **Обучающая выборка (training set)** – конечный набор обьектов, для которых известны значения целевой переменной

Пример: набор ресторанов, открытых более года назад, для которых известна их прибыль За первый год

Признаки описывают обьекты с помощью чисел

Специалист по анализу данных не является экспертом в предметной области – вся необходимая информация содержится в обучающей выборке. Эксперты нужны при формировании признаков.

Стандартная постановка задачи



Обучение алгоритма

- Есть обучающая выборка и функционал качества
- Семейство алгоритмов $\mathcal{A}$
    - Из чего выбираем алгоритм
    - Пример: все линейные модели
    - $\mathcal{A} = w0 + w1x1 + \cdots + wdxd\ w0, w1, \dots, wd \in \mathbb{R}\}$
- Обучение: поиск оптимального алгоритма с точки зрения функционала качества

Виды признаков

- Числовые
- Бинарные
- Категориальные (название города, марка машины)
- Признаки со сложной внутренней структурой (изображение)

Виды данных

- Таблицы
- Текстовые данные
- Изображения
- Звук
- Логи

Большинство алгоритмов машинного обучения работает с числовыми данными,
Поэтому все виды данных необходимо переводить в числа

Типы задач в зависимости от целевой переменной

Классификация

- $Y = \{0,\ 1\}$ – классификация на 2 класса
- $Y = \{1, \dots, M\}$– классификация на M непересекающихся классов
- $Y = \{0, 1\}^M$ - классификация на M классов, которые могут пересекаться

Регрессия

- $Y = R$
- $Y = R^n$

Ранжирование

- $\boldsymbol{Y}$ – конечное упорядоченное множество

Задачи без целевой переменной

- Кластеризация – задача разделения обьектов на группы,
  при этом где целевые переменные для обьектов неизвестны
(или не существуют). Разделение происходит только на основе
 признаковых описаний обьектов

- Понижение размерности – задачи генерации новых признаков
  (по числу меньше, чем старых), так что с помощью их задача
  решается не хуже чем с исходными

- Оценивание плотности – задача приближения распределения обьектов

- Визуализация – задача изображения многомерных обьектов
  в 2х или 3х мерном пространстве с сохранением зависимостей
  между ними

Обучение с учителем

- Если нам известны значения целевой переменной, то есть алгоритм обучается так, чтобы правильно предсказывать целевую переменную – это обучение с учителем :
    - Классификация
    - Регрессия
    - Ранжирования

Обучение без учителя

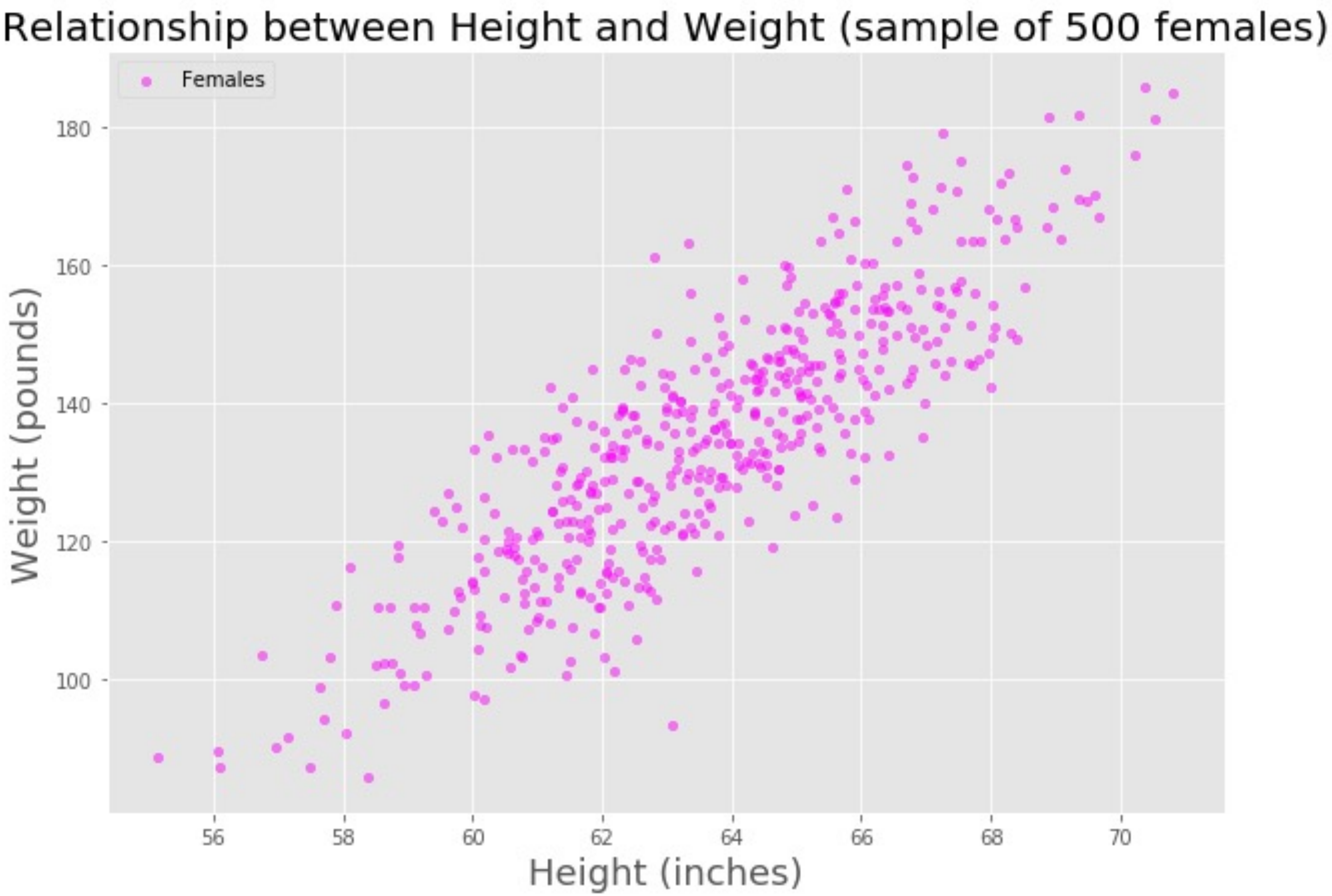
- Если нам неизвестны значения целевой переменной или целевая переменная вообще отсутствует, то есть алгоритм обучается только по признакам обьектов, то это обучение без учителя. Примерами обучения с учителем являются кластеризация, понижение размерности и др.

Алгоритм решения задачи

1. Постановка задачи
2. Выделение признаков
3. Формирование выборки
4. Выбор функции потерь и метрики качества
5. Предобработка данных
6. Построение модели
7. Оценивание качества модели

Линейная регрессия

Парная регрессия

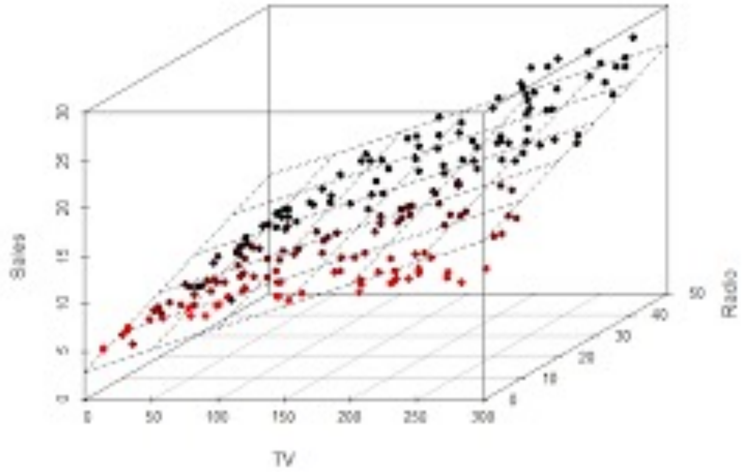


Парная регрессия

- Простейший случай: один признак
- Модель: $a(x) = w_1 x + w_0$
- Два параметра w_1, w_0
- w_1 - тангенс угла наклона
- w_0 - где прямая пересекает ось ординат

Два признака

- Случай посложнее: два признака
- Модель: $a(x) = w_2 x_2 + w_1 x + w_0$
- Три параметра w_1, w_0, w_2



Много признаков

- Общий случай: d признаков
- количество параметров: $d + 1$
- Модель: $a(x) = w_0 + w_1 x_1 + \ldots + w_d x_d$

Веса/коэффиценты

Свободный коэффицент/сдвиг/bias

Много признаков

$a(x) = w_0 + w_1 x_1 + \ldots + w_d x_d = w_0 + < w, x >$

Будем считать, что есть признак, всегда равный единице:

$a(x) = w_0 + w_1 x_1 + \ldots + w_d x_d = w_1 * 1 + w_2 x_2 + \ldots + w_d x_d = < w, x >$

Применимость линейной регрессии

$a(x) = w_0 + w_1 x_1 + \ldots + w_d x_d = w_0 + < w, x >$

- Нет гарантий, что целевая переменная именно так зависит от признаков
- Надо формировать признаки так, чтобы модель подходила

Предсказание стоимости квартиры

$a(x) = w_0 + w_1 *$ (площадь) + w_2 * (район) + w_3 * (расстояние до метро)

- Признаки: площадь, район, расстояние до метро
- Целевая переменная: рыночная стоимость квартиры

Предсказание стоимости квартиры

$a(x) = w_0$ + $\boldsymbol{w_1}$* **(площадь)**

w_2 * (район) +

w_3 * (расстояние до метро)

- За каждый квадратный метр добавляем к прогнозу

Предсказание стоимости квартиры

$a(x) = w_0$ + w_1* (площадь)

w_2 * (район) +

w_3 * (расстояние до метро)

- Что-то странное

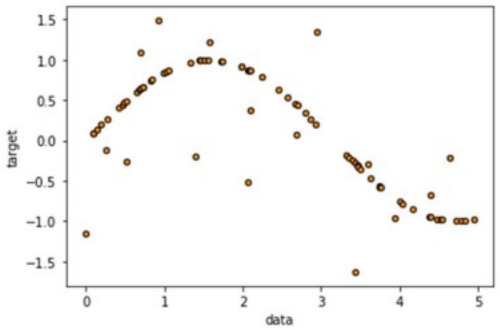
Предсказание стоимости квартиры

$a(x) = w_0$ + w_1* (площадь)

w_2 * (район) +

w_3 * (расстояние до метро)

- Что-то странное



Кодирование категориальных признаков

- Значение признака район: $U = \{u_1, \dots, u_m\}$
- Новые признаки вместо x_j: $[x_j = u_1], \dots, [x_j = u_m]$
- One-hot кодирование

Район
ЦАО
ЮАО
ЦАО
САО
ЮАО

ЦАО	ЮАО	САО
1	0	0
0	1	0
1	0	0
0	0	1
0	1	0

Кодирование категориальных признаков

$a(x) = w_0$ + w_1* (площадь)

w_2 * (квартира в ЦАО?) +

w_2 * (квартира в ЮАО?)) +

w_2 * (квартира в САО?)) +

w_3 * (расстояние до метро)

Район
ЦАО
ЮАО
ЦАО
САО
ЮАО

→

ЦАО	ЮАО	САО
1	0	0
0	1	0
1	0	0
0	0	1
0	1	0

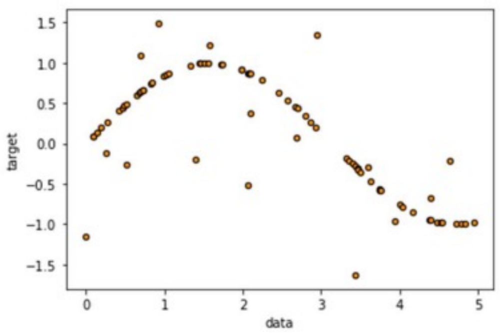
Предсказание стоимости квартиры

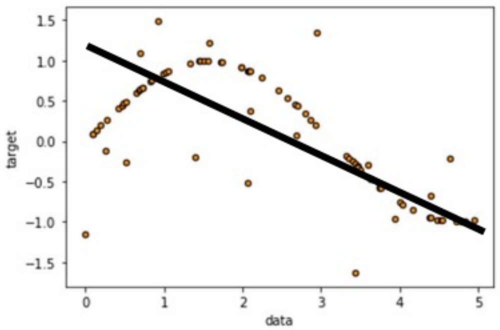
$a(x) = w_0$ + w_1* (площадь)

w_2 * (район) +

w_3 * (расстояние до метро)

- Что-то странное





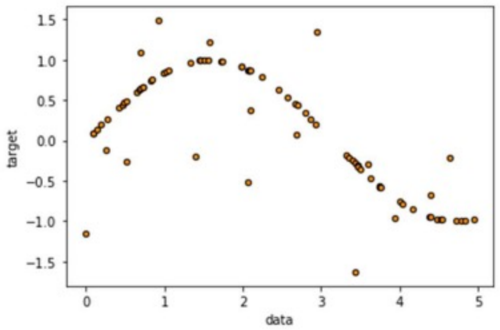
Предсказание стоимости квартиры

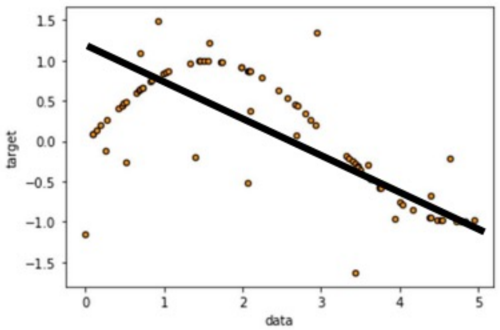
$a(x) = w_0$ + w_1* (площадь)

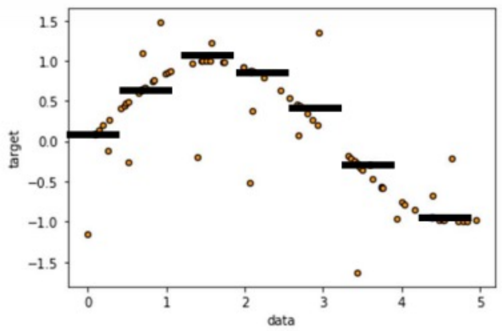
w_2 * (район) +

w_3 * (расстояние до метро)

- Что-то странное





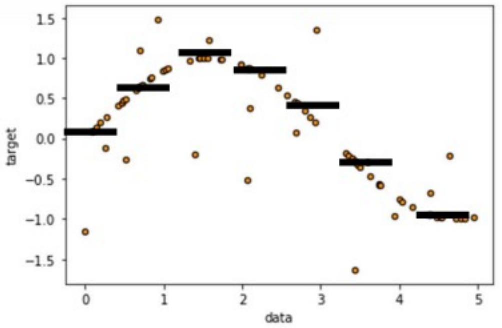


Предсказание стоимости квартиры

$a(x) = w_0$ + w_1* (площадь)

w_2 * (район) +

$\boldsymbol{w_3} * [\mathbf{t_0} \leq \mathbf{x_3} < \mathbf{t_1}] + \ \ldots + \mathbf{w_{3+n}}[\mathbf{t_{n-1}} \leq \mathbf{x_3} < \mathbf{t_n}]$



Линейные модели

- Модель линейной регрессии хороша, если признаки сделаны специально под нее

- Пример: one-hot кодирование категориальных признаков или бинаризация числовых признаков

Валидация моделей

Функция потерь для регрессии

- Частый выбор – квадратичная функция потерь
  $L(y, a) = (a - y)^2$

- Функционал ошибки – среднеквадратичная ошибка(mean squered error, MSE)

$$Q(a, X) = \frac{1}{l}\sum_{i=1}^{l}(a(x_i) - y_i)^2$$

- Штрафует за выбросы

Функция потерь для регрессии

- Еще один вариант – средняя абсолютная ошибка (mean absolute error, MAE)

$$Q(a, X) = \frac{1}{l} \sum_{i=1}^{l} |a(x_i) - y_i|$$

- Слабее штрафует за выбросы
- Зануляет признаки перед незначищими признаками

Линейная регрессия в векторном виде

$$a(x) = < w, x >$$

- Среднеквадратичная ошибка и задача обучения:

$$\frac{1}{l}\sum_{i=1}^{l}(< w, x_i > - y_i)^2 \rightarrow \min\_w$$

Матрицы

- Матрица – таблица с числами
- Матрица ”обьекты-признаки” :

Обьект и его признаки

$$X = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1d} \\ \boxed{x_{21} \quad x_{22} \quad \cdots \quad x_{2d}} \\ \vdots & \vdots & \ddots & \vdots \\ x_{\ell 1} & x_{\ell 2} & \cdots & x_{\ell d} \end{pmatrix} \in \mathbb{R}^{\ell \times d}$$

Матрицы

- Матрица – таблица с числами
- Матрица ”обьекты-признаки” :

Обьект и его признаки

$$X = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1d} \\ x_{21} & x_{22} & \cdots & x_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ x_{\ell 1} & x_{\ell 2} & \cdots & x_{\ell d} \end{pmatrix} \in \mathbb{R}^{\ell \times d}$$

$$X = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1d} \\ x_{21} & x_{22} & \cdots & x_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ x_{\ell 1} & x_{\ell 2} & \cdots & x_{\ell d} \end{pmatrix} \in \mathbb{R}^{\ell \times d}$$

Признак на всех обьектах

Векторы

- Вектора размер d – тоже матрица размера $1 \ * \ d$
- Вектор-строка: $w = (w_1, \dots, w_d) \in \ R^{1 \ *d}$
- Вектор-столбец: $w = (w_1, \dots, w_d)^T \in \ R^{d \ *1}$

Применение линейной модели

- $a(x) = < w, x > = w_1 x_1 + \ ... + w_d x_d$

- Как применить модель к обучающей выборке?

$$\begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1d} \\ x_{21} & x_{22} & \cdots & x_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ x_{\ell 1} & x_{\ell 2} & \cdots & x_{\ell d} \end{pmatrix}$$

$$\begin{pmatrix} w_1 \\ w_2 \\ \vdots \\ w_d \end{pmatrix}$$

$$\begin{pmatrix} \sum_{i=1}^{d} w_i x_{1i} \\ \sum_{i=1}^{d} w_i x_{2i} \\ \vdots \\ \sum_{i=1}^{d} w_i x_{\ell i} \end{pmatrix}$$

Матричное умножение

- Только для матриц $A \in R^{m*k}$ и $\mathrm{B} \in R^{k*n}$
- Результат: и $AB = C \in R^{m*n}$
- $\mathrm{c}_{ij} = \sum_{p=1}^{k} a_{ip} b_{pj}$

Пример

$$\begin{pmatrix} 1 & 2 \\ 0 & 1 \\ 10 & 0 \end{pmatrix} * \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 2 \end{pmatrix} = \begin{pmatrix} & & \\ & & \\ & & \end{pmatrix}$$

$$\begin{pmatrix} \boxed{1 \quad 2} \\ 0 \quad 1 \\ 10 \quad 0 \end{pmatrix} * \begin{pmatrix} \boxed{\begin{matrix} 1 \\ 0 \end{matrix}} & \begin{matrix} 0 \\ 0 \end{matrix} & \begin{matrix} 0 \\ 2 \end{matrix} \end{pmatrix} = \begin{pmatrix} \boxed{1} & & \\ & & \\ & & \end{pmatrix}$$

$$\begin{pmatrix} \boxed{1 \quad 2} \\ 0 \quad 1 \\ 10 \quad 0 \end{pmatrix} * \begin{pmatrix} \begin{matrix} 1 \\ 0 \end{matrix} & \boxed{\begin{matrix} 0 \\ 0 \end{matrix}} & \begin{matrix} 0 \\ 2 \end{matrix} \end{pmatrix} = \begin{pmatrix} 1 & \boxed{0} & \\ & & \\ & & \end{pmatrix}$$

$$\begin{pmatrix} \boxed{1 \quad 2} \\ 0 \quad 1 \\ 10 \quad 0 \end{pmatrix} * \begin{pmatrix} \begin{matrix} 1 \\ 0 \end{matrix} & \begin{matrix} 0 \\ 0 \end{matrix} & \boxed{\begin{matrix} 0 \\ 2 \end{matrix}} \end{pmatrix} = \begin{pmatrix} 1 & 0 & \boxed{4} \\ & & \\ & & \end{pmatrix}$$

Модель линейной регрессии

Среднеквадратичная ошибка и задача обучения:

$$\frac{1}{l}\sum_{i=1}^{l}(< w, x_i > \quad -y_i)^2 \to \min\_w$$

Отклонения прогнозов от ответов:

$$Xw - y = \begin{pmatrix} \langle w, x_1 \rangle - y_1 \\ \vdots \\ \langle w, x_\ell \rangle - y_\ell \end{pmatrix}$$

Вычисление ошибки

Евклидова норма:

$$||z|| = \sqrt{\sum_{j=1}^{n} z_j^2}$$

$$||z||^2 = \sum_{j=1}^{n} z_j^2$$

$$Xw - y = \begin{pmatrix} \langle w, x_1 \rangle - y_1 \\ \vdots \\ \langle w, x_\ell \rangle - y_\ell \end{pmatrix}$$

Вычисление ошибки

Отклонения прогнозов от ответов:

$$||z|| = \sqrt{\sum_{j=1}^{n} z_j^2}$$

Среднеквадратичная ошибка:

$$\frac{1}{l}||Xw - y||^2 = \frac{1}{l}\sum_{i=1}^{l}(< w, x_i > - y_i)^2$$

Обучение линейной регрессии

$$\frac{1}{l}||Xw - y||^2 \to \min\_w$$

Обобщающая способность

Как готовиться к экзамену?

Заучить все примеры с занятий

Переобучение (overfitting)

Хорошее качество на обучении
Низкое качество на новых данных

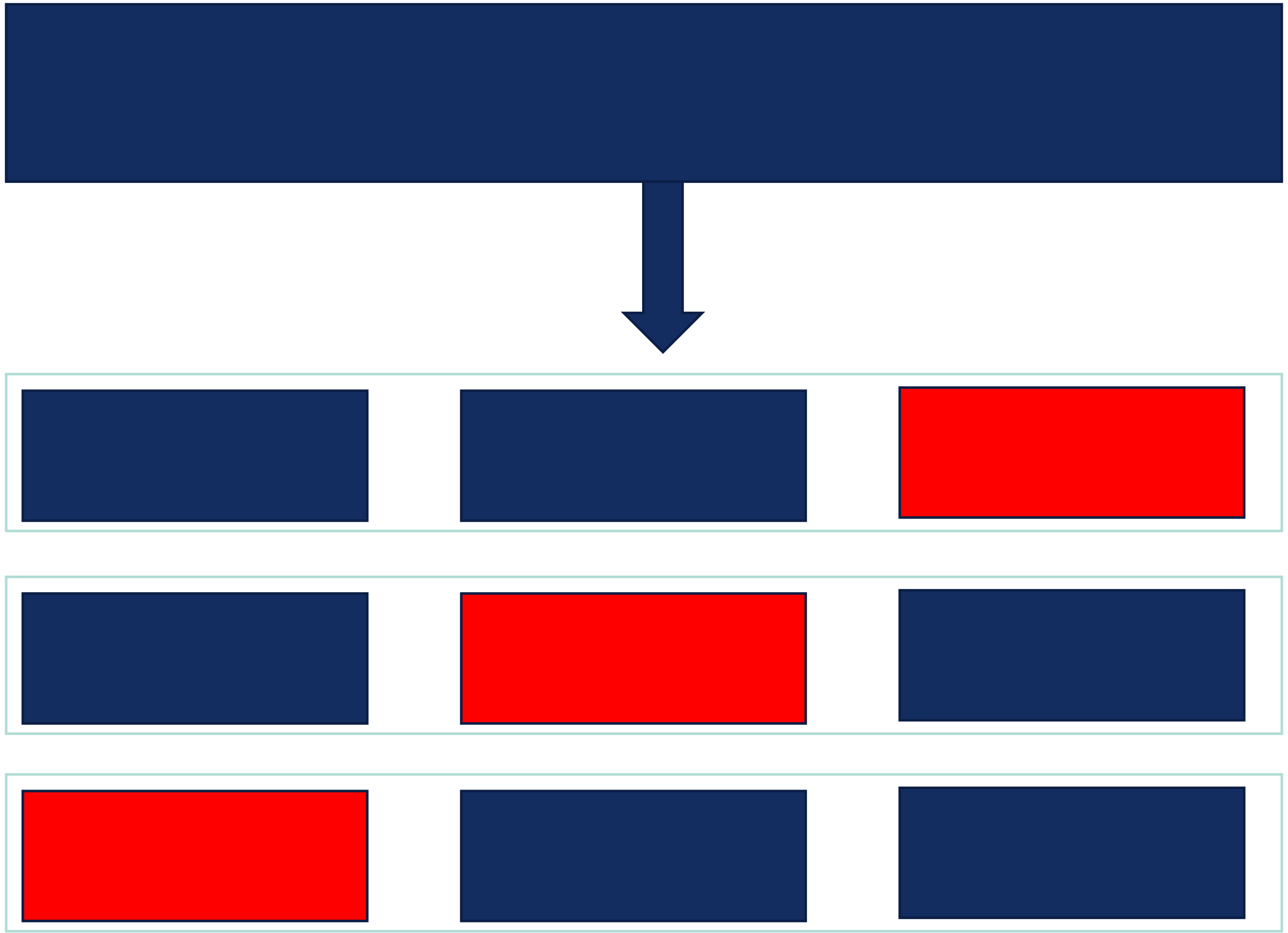
Разобраться в предмете и усвоить алгоритмы решения задач

Обобщение (generalization)

Отложенная выборка

- Слишком большое обучение — тестовая выборка нерепрезентативна
- Слишком большой тест — модель не сможет обучиться
- Обычно: 70/30, 80/20

Кросс-валидация



Кросс-валидация

- Надёжнее отложенной выборки, но медленнее
- Параметр — количество разбиений n (фолдов, folds)
- Хороший, но медленный вариант — n = l(leave-one-out)
- Обычно: n = 3 или n = 5 или n = 10

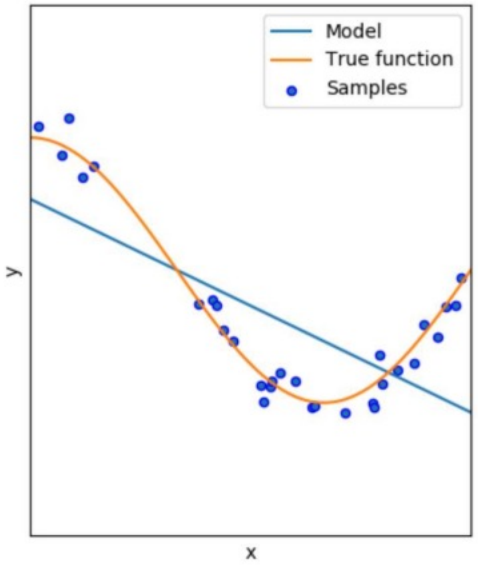
Признаки переобученной модели

- Большая разница в качестве на тренировочных и
  и тестовых данных(модель подгоняется под тренировочные
  данные и не может найти истинная зависимость)

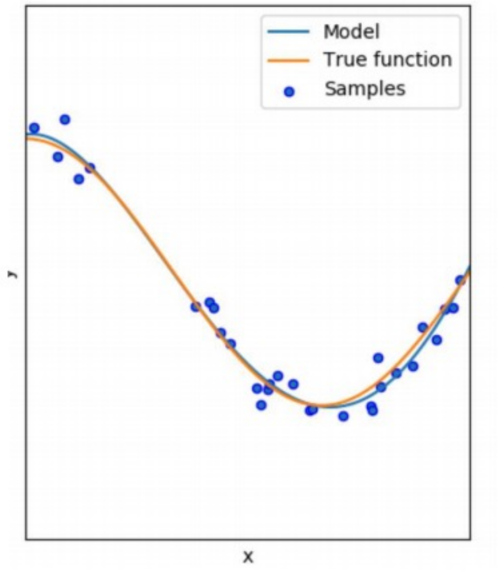
- Большие значения параметров (весов) w_j модели

Нелинейная задача

$$a(x) = w_0 + w_1 x$$

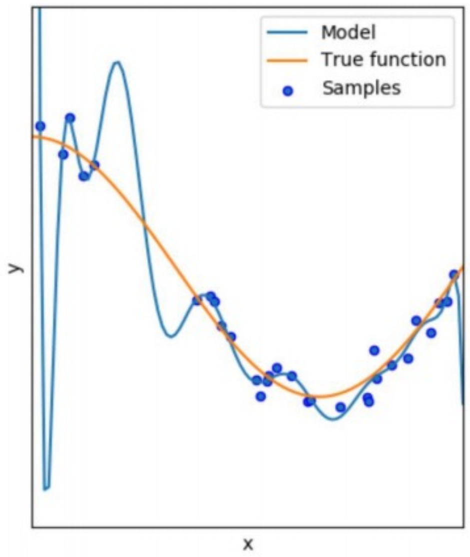


$$a(x) = w_0 + w_1 x + w_2 x^2 + w_3 x^3 + w_4 x^4$$



Нелинейная задача

$$a(x) = w_0 + w_1 x + w_2 x^2 + w_3 x^3 + w_4 x^4 + \cdots + w_{15} x^{15}$$



Симптом переобучения

$$a(x) = 0.5 + 134545262x \; - 345345346x^2 + \; \dots.$$

- Большие коэффиценты – симптом переобучения
- Эмпирическое наблюдение
- Пример: предсказание роста по весу
- Изменение веса на 0.01 кг приведет к изменению роста на 7 см
- Не похоже не правильную зависимость

Регуляризация

- Будем штрафовать за большие веса!
- Пример функционала:

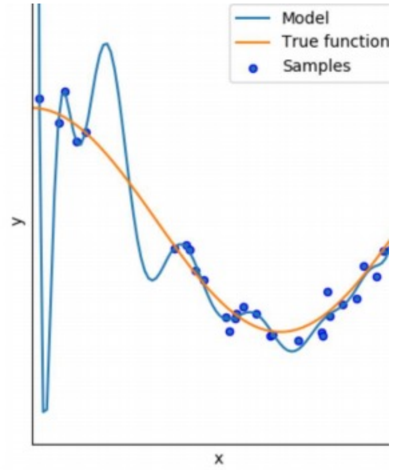
$$Q(a, X) = \frac{1}{l}\sum_{i=1}^{l}(< w, x_i > \; - y_i)^2$$

- Регуляризатор:

$$||w||^2 = \sum_{j=1}^{d} w_j^2$$

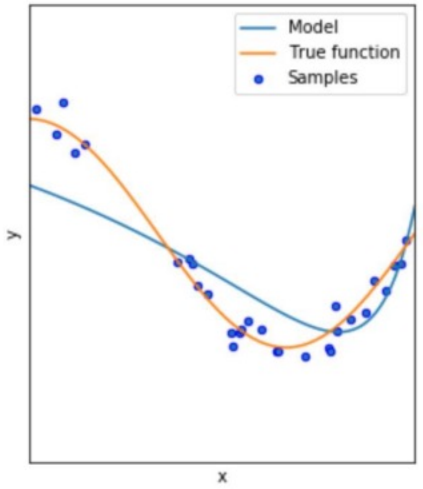
Эффект регуляризации

- $a(x) = w_0 + w_1 x + w_2 x^2 + \ \dots + w_{15} x^{15}$
- $\frac{1}{l}\sum_{i=1}^{l}(a(x_i) - y_i)^2 \to \min\_w$



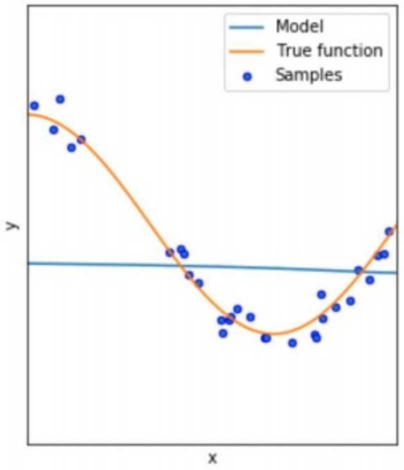
Эффект регуляризации

- $a(x) = w_0 + w_1 x + w_2 x^2 + \ \dots + w_{15} x^{15}$
- $\frac{1}{l}\sum_{i=1}^{l}(a(x_i) - y_i)^2 + 1\left||w|\right|^2 \to \min\_w$



Эффект регуляризации

- $a(x) = w_0 + w_1 x + w_2 x^2 + \ \ldots + w_{15} x^{15}$
- $\frac{1}{l} \sum_{i=1}^{l} (a(x_i) - y_i)^2 + 100 \left|\left|w\right|\right|^2 \to \min\_w$



Лассо

- Регуляризованный функионал
- $\frac{1}{l}\sum_{i=1}^{l}(< w, x_i > -y_i)^2 + \lambda||w_j||^2 \to \min\_w$
- LASSO(Least Absolute Shrinkage and Selection Operator)
- Некоторые веса зануляются
- Приводит к отбору признаков

Регуляризаторы

- $||z||_2 = \sqrt{\sum_{j=1}^{d} z_j^2}$ – L_2 – норма
- $||z||_1 = \sum_{j=1}^{d} |z_j|$ – L_1 – норма

Гиперпараметры

- Гиперпараметр алгоритма — то, что задается вручную
- Пример: коэффициент регуляризации
- Параметр алгоритма — то, что определяется моделью
- Пример: веса регрессии
- Как подбирать гиперпараметры алгоритма?
- Нельзя подбирать по обучающей выборке — это приведет к переобучению
- Нужно использовать дополнительные данные (валидация)

Чуть больше терминов

- После подбора всех гиперпараметров стоит проверить на совсем новых данных, что модель работает
- Обучающая выборка — построение модели
- Валидационная выборка — подбор гиперпараметров модели
- Тестовая выборка — финальная оценка качества модели

Предсказание стоимости квартиры

$$a(x) = 100.000 \ * (\text{площадь})$$
$$+500.000 \ * (\text{число магазинов рядом})$$
$$+100 * (\text{средний доход жильцов дома})$$

Чем больше вес, тем важнее признак?

Только если признаки масштабированы!

Масштабирование признаков

- Отмасштабируем j-й признак
- Вычисляем среднее и стандартное отклонение признака на обучающей выборке:

$$\mu_j = \frac{1}{l}\sum_{i=1}^{l} x_i^j$$

$$\sigma_j = \sqrt{\frac{1}{l}\sum_{i=1}^{l}\left(x_i^j - \mu_j\right)^2}$$

Масштабирование признаков

- Вычтем из каждого значения признака среднее и поделим на стандартное отклонение:

$$x_i^j \coloneqq (x_i^j - \mu_j) / \sigma_j$$

Регуляризация

- Если модель переобучается, то веса используются для запоминания обучающей выборки
- Правильнее масштабировать признаки и регуляризовать модель перед изучением весов

Пример

- 1000 объектов • Два признака
- Первый принимает значения от 0 до 1
- Второй равен единице на 10 объектах и нулю на 990 объектах
- $y = x_1 + 2x_2$
- Удаляем первый признак, получаем MSE = 0.08
- Удаляем второй признак, получаем MSE = 0.04
- Правильнее удалить признак и посмотреть, как сильно растет ошибка без него

Пример

```
[0.3175037 , 1.        ],
[0.59558502, 1.        ],
[0.48660609, 1.        ],
[0.69255463, 1.        ],
[0.81968981, 1.        ],
[0.48844247, 1.        ],
[0.13426702, 1.        ],
[0.850628  , 1.        ],
[0.57499033, 1.        ],
[0.73993748, 1.        ],
[0.70466465, 0.        ],
[0.96821177, 0.        ],
[0.29530732, 0.        ],
[0.70530677, 0.        ],
[0.36567633, 0.        ],
[0.39541072, 0.        ],
[0.23059464, 0.        ],
[0.34401018, 0.        ],
[0.94829675, 0.        ],
[0.29257085, 0.        ],
[0.24599061, 0.        ],
[0.58313798, 0.        ],
```

Спасибо за внимание!

Ildar Safilo

@Ildar_Saf
irsafilo@gmail.com
https://www.linkedin.com/in/isafilo/